Minimum Regret Approach to Network Management under Uncertainty with Application to Connection Admission Control and Routing♣

Vladimir Marbukh

National Institute of Standards and Technology
100 Bureau Drive, Stop 8920
Gaithersburg, MD 20899-8920, USA
marbukh@nist.gov

Abstract. This paper proposes a framework for network management intended to balance network performance under normal steady operational conditions with robustness under non-steady, and/or adverse conditions. Working in conjunction with anomaly and/or intrusion detection, the proposed framework allows the network to develop a set of measured responses to possible anomalies and external threats by minimizing the average maximum network losses, i.e., regrets or risks, due to uncertainty. Loss maximization guards against uncertainty within each scenario. Averaging of the maximum losses reflects any available information on the likelihood of different scenarios. The proposed framework includes Bayesian and minimax approaches as particular cases. The paper demonstrates how the proposed framework can alleviate high sensitivity of a cost-based admission and routing scheme to uncertain resource costs and possible presence of excessive and/or adversarial traffic. Specific examples include competitive and minimum interference admission and routing schemes.

1 Introduction

Different approaches to providing Quality of Service (QoS) guarantees by a system of shared resources include using Traffic Controlling Devices (TCD) for traffic shaping and policing, Fair Traffic Scheduling Mechanisms (TSM), and Measurement Based Control (MBC). TCD and TSM allow the network to provide QoS guarantees to a source complying with its service agreement by making admission control decisions based on the worst case scenario. However, due to a wide range of traffic patterns generated by current and especially future applications, this approach runs a risk of significant over provisioning of the network resources under normal operating conditions. This risk can be mitigated by MBC, which extracts control

♣ This work was supported in part by DARPA under NMS program.

information from on-line measurements [1]. This real-time control information can be used by the resources to control usage and/or by users to modify their behavior according to the smart market concept [2]-[7]. Reliable extraction of the control information from the measurements and optimal behavior of smart market are achievable in equilibrium. However, even in equilibrium MBC faces a problem of making decisions under uncertainty due to the statistical nature of measurements and state aggregation needed to reduce amount of signaling information. The uncertainty becomes a critical problem for a network operating under non-steady and/or adversarial conditions. These conditions may be a result of natural events, i.e., fiber cuts, flash crowds, and/or adversary attack, i.e., denial of service attacks [8]. Since anomaly and intrusion detection procedures are based on statistical inferences, such as hypotheses testing, clustering, discriminant analysis, etc., the result of the detection is typically a set of possible scenarios and their likelihood. Uncertainty within a scenario may be a result of interval rather than point estimates for the state of environment specified by the scenario. For example, for flash crowds and distributed denial of service attacks a scenario represents a subset of flows - an aggregate, responsible for congestion [8]. Uncertainty results from difficulty to identify the aggregates and inability to distinguish between legitimate and excessive traffic within each aggregate.

Currently commercial networks including the Internet may carry mission-critical applications with wide range of bandwidth and QoS requirements. Possibility of anomalies caused by sudden increase in offered traffic, fiber cut and/or an adversary attack necessitates developing network management schemes that balance cost efficiency under normal operating conditions with robustness under non-steady and/or adversarial conditions. The right balance between cost efficiency and performance can be achieved by adaptive sharing of the network management responsibilities between MBC, TCD and TSM. The critical role in this collaboration belongs to MBC which determines the optimal "weight" of each scheme by adjusting parameters of the TCD and/or TSM algorithms.

Section 2 of this paper proposes a framework for network management under uncertainty. Working in conjunction with anomaly and/or intrusion detection, the proposed framework allows the network to develop a set of measured responses to possible anomalies and external threats by minimizing the average maximum network losses, i.e., regrets or risks, due to uncertainty. Loss maximization guards against uncertainty within each scenario. Averaging of the maximum losses reflects an available information on the likelihood of different scenarios. The proposed framework includes Bayesian and minimax approaches as particular cases. Section 3 demonstrates how the proposed framework can alleviate high sensitivity of cost-based admission and routing strategies to the uncertain resource costs and possible presence of excessive and/or adversarial traffic. Section 4 considers specific examples of competitive [9] and minimum interference admission and routing schemes [10].

2 Risk Assessment and Management

Throughout the paper we assume that the network utility $W(u|\theta)$, typically the revenue generated by the network, is a known function of the network control action u and environment θ. Vector u may describe pricing, resource allocation, admission control, routing, scheduling etc., and vector θ may characterize topology, resource capacities, traffic sources, etc. We consider a problem of selecting the network control action u under incomplete information on the state of environment θ. This section discusses approaches to assessment and minimization of the risks due to uncertainty.

2.1 Risk Assessment

If the state of environment θ is known, the optimal control action $u = u^*(\theta)$ maximizes the network utility $W(u|\theta)$:

$$u^*(\theta) = \arg\max_{u \in U} W(u|\theta) \ . \quad \textbf{(1)}$$

Formula (1) determines the best network response to given θ. In presence of other players besides the network, e.g., users, each player determines his best response to the other player strategies and the state of environment θ by maximizing his utility function. Harsanyi transformation of sequential games with incomplete information [11] assumes that the state of environment θ is selected by "Nature" at the beginning of the game according to some probability distribution $p_0(\theta)$ known to all players. Players may have private information regarding selected vector θ and they update their information on the selected vector θ according to Bayesian rules as game progresses. According to this approach, the best network response is

$$u^* = \arg\max_{u \in U} \sum_{\theta} W(u,\theta)\, p(\theta) \quad \textbf{(2)}$$

where $p(\theta)$ is the updated distribution $p_0(\theta)$. According to the theory of non-cooperative games, once the best responses by players are determined, reasonable players should select their equilibrium strategies. Note that a number of equilibrium concepts exist, including the most widely used Nash equilibrium.

In this paper we are interested in a case when the "Nature", or as we prefer to call it "environment" may have malicious intend directed towards some players, i.e., the network and some users. Due to space constraints, in this paper we only consider a case when this malicious intend may be directed towards the network. Following

Savage [12] we characterize the network loss in performance resulted from non-optimal selection of the control action u due to uncertain environment θ by the following regret or loss function:

$$L(u|\theta) = \max_{u' \in U} W(u'|\theta) - W(u|\theta) \ . \quad \textbf{(3)}$$

Local maximums $\theta_i^* = \theta_i^*(u)$ of the loss function (3) over θ represent different risk factors for the network due to uncertain environment (for example see [13]). We propose to model uncertain environment as a player with utility function (3) and certain restrictions on the set of strategies $\theta \in \Theta$. This assumption leads to a framework for network management under uncertainty, which, in effect, is a systematic approach to balancing different risk factors. Bayesian and minimax approaches are particular cases of this framework.

Note that there is a certain degree of freedom in selecting loss function $L(u|\theta)$. This selection reflects the desired balance between different risk factors. Using loss function (3) in networking context has been proposed in [13]. Competitive approach to admission control and routing [9] guards the network against the worst case scenario with respect to the state of environment θ by containing the losses

$$L(u|\theta) = \frac{1}{W(u|\theta)} \max_{u' \in U} W(u'|\theta) - 1 \quad \textbf{(4)}$$

where vector of control action u represents admission control and routing, and vector θ represents completely unknown sequence of future request arrivals. The competitive approach assumes that the holding times are known upon request arrival, at least in probabilistic sense. Another basic assumption is that exact information on the instantaneous link utilization is available.

2.2 Risk Management

Network operating under non-steady and/or adversarial conditions cannot rely on extensive historical data to estimate the state of environment θ. However, some aggregated information on θ is often available. We assume that this information can be quantified in terms of probabilities $p_i = \Pr ob(\theta \in \Theta_i)$, $\sum_i p_i = 1$ for some partition of the region of all possible vectors $\Theta = \{\theta\}$ into set of mutually exclusive regions Θ_i: $\bigcup_i \Theta_i = \Theta$, $\Theta_i \bigcap \Theta_j = \varnothing$. Given scenario $\{p_i, \Theta_i\}$, the optimal network response is

$$u^* = \arg\min_{u \in U} \hat{L}(u) \tag{5}$$

where the performance loss (regret or risk) due to uncertainty is

$$\hat{L}(u) = \sum_i p_i \max_{\theta \in \Theta_i} L(u|\theta) \ . \tag{6}$$

Procedure (5)-(6) can be interpreted as a game between the network and environment with constraints on the set of environment strategies. In an extreme case when each subset Θ_i is a singleton, i.e., consists of a single point θ, procedure (5)-(6) reduces to the Bayesian procedure (2). In another extreme case when Θ is partitioned into itself and the empty set $\varnothing$, procedure (5)-(6) can be interpreted as a zero sum game between the network and environment. Since anomaly and/or intrusion detection typically results in a set of possible scenarios $s = \{1,..,S\}$ rather than one scenario, the network faces a task of balancing risks associated with different scenarios. This task can be formulated a multi criteria optimization problem

$$\min_{u \in U} \left(R_1(u),..,R_S(u) \right) \tag{7}$$

where risk associated with scenario s is $R_s(u) = \hat{L}_s(u) - \min_{u' \in U} \hat{L}_s(u')$, and the loss in performance for scenario s is $\hat{L}_s(u)$. Pareto frontier yields a reasonable set of the network control actions $u \in U$ in this situation. Any additional information can be used to reduce the set of Pareto solutions. For example, if anomaly and/or intrusion detection can identify (subjective) probabilities π_s of different scenarios s, $\sum_s \pi_s = 1$, then the network may minimize the average (Bayesian) risk

$$\min_{u \in U} \sum_s \pi_s R_s(u) \ . \tag{8}$$

In a case of unknown π_s and/or adversarial environment the network may prefer to minimize the maximum risk

$$\min_{u \in U} \max_{s=1,..,S} R_s(u) \ . \tag{9}$$

Working in conjunction with anomaly and/or intrusion detection the network should perform two tasks. The long time scale task is maintaining and updating a library of possible scenarios s and subroutines for calculating the corresponding risks $R_s(u)$. The short time scale task is risk minimization (7).

3 Connection Admission Control and Routing under Uncertainty

Admission of a request, on the one hand, brings revenue to the network, but, on the other hand, ties up the network resources until the service is completed, and thus may cause future revenue losses due to insufficient resources for servicing future requests. The implied cost of a resource represents this potential revenue loss, and the surplus value is the difference between the revenue brought by the admitted request and the implied cost of the occupied resources. An incoming request should be accepted if the surplus value is positive, and should be rejected otherwise. This section demonstrates how proposed approach to risk assessment and management can be used to balance the performance and robustness of a cost based connection admission control and routing scheme under uncertainty.

3.1 Risk Associated with Connection Admission and Routing

Consider an arriving request for bandwidth b on a route $r \in \{r_1,..,r_k\}$, where the set of feasible routes $\{r_1,..,r_k\}$ is determined by the origin-destination of the request, availability of the bandwidth, maximum allowed number of hops, QoS requirements, etc. Let the implied cost of tying up bandwidth b along route r be c_r where we dropped b from notations. Given the set of feasible routes and route costs, the optimal route r_* and admission condition for an arriving request willing to pay rate w are as follows:

$$c_* \equiv c_{r_*} = \min_{r\in\{r_1,..,r_k\}} c_r \le w \ . \qquad \textbf{(10)}$$

The implied costs c_r are determined by future events such as future request arrivals, holding times, availability of resources, topology, etc. Inability of the network to predict these future events, especially in non-steady and/or adversarial environment, is a source of uncertainty in the implied costs c_r. Uncertainty in w may be caused by presence of excessive and/or malicious traffic such as in flash crowds or denial of service attack [8]. Utility of the admission and routing decisions can be characterized by the surplus value

$$W(r|w,c) = \begin{cases} w - c_r & if\ r \neq \varnothing \\ 0 & if\ r = \varnothing \end{cases} \qquad \textbf{(11)}$$

where the request is accepted on route r if $r \neq \varnothing$, and is rejected if $r = \varnothing$. For utility function (11) the loss function (3) takes the following form:

$$L(r|w,c)=\begin{cases}\max\{0,w-c_*\}+c_r-w & if\ r\neq\varnothing\\ \max\{0,w-c_*\} & if\ r=\varnothing\end{cases}. \quad \textbf{(12)}$$

3.2 Approximation of Separable Route Costs

Computational feasibility to minimize losses (12) over r critically depends on the range Θ of possible vectors of implied costs $(c_r : r = r_1,..,r_k)$. In this subsection we consider a case of separable route costs:

$$\Theta = \underset{r=r_1,..,r_k}{\otimes} [\breve{c}_r, \hat{c}_r]. \quad \textbf{(13)}$$

Note that (13) is a "first order" approximation to more realistic scenarios since implied costs c_r of different routes strongly correlate with each other due to the global nature of implied costs and route overlapping.

In a symmetric case: $c_r \in [\breve{c}, \hat{c}]$, $\forall r = r_1,..,r_k$, the network has two pure strategies: reject the request, and accept the request on a randomly selected route $r = r_1,..,r_k$. The adversarial environment has two pure strategies $c_r = \breve{c}$ and $c_r \in \hat{c}$ for $r = r_1,..,r_k$. The corresponding payoff matrix (12) is:

$$\begin{array}{lll} & reject: & accept: \\ \breve{c}: & \max\{0, w-\breve{c}\} & \max\{0, w-\breve{c}\}+\breve{c}-w. \\ \hat{c}: & \max\{0, w-\hat{c}\} & \max\{0, w-\hat{c}\}+\hat{c}-w \end{array} \quad \textbf{(14)}$$

Game (14) has different solutions in the following three cases. In cases $w \le \breve{c}$ and $w \ge \hat{c}$ game (14) has the saddle point and the network has pure optimal strategy to reject and, respectively, accept the request. In a case $\breve{c} < w < \hat{c}$ game (14) does not have saddle point and optimal strategy for the network is mixed: reject the request with probability $1-\alpha$, and accept with probability α, where $\alpha = (w-\breve{c})/(\hat{c}-\breve{c})$.

In a general case of separable route costs (13) if $w \le \breve{c}_* \equiv \min_r \breve{c}_r$ or $w \ge \hat{c}_* \equiv \min_r \hat{c}_r$, then the corresponding game has the saddle point and the network has pure optimal strategy to reject and, respectively, accept the request on a route r^{opt}: $\hat{c}_{r^{opt}} \equiv \min_r \hat{c}_r$. In a case $\breve{c}_* < w < \hat{c}_*$ the corresponding game does not have the saddle point and the optimal strategy is closely approximated by the following mixed strategy: reject the request with probability $1-\alpha$, and accept with

probability α on route r^{opt}, where α and r^{opt} are determined by the following optimization problem:

$$\alpha = \max_{r=r_1,..,r_k} \left\{ \frac{w - \breve{c}_r}{\hat{c}_r - \breve{c}_r} \right\}. \quad \textbf{(15)}$$

4 Examples

This section presents examples, including an aggregated version of the competitive scheme under uncertain holding times, and an aggregated version of the Minimum Interference Routing Algorithm (MIRA).

4.1 Aggregated Competitive Scheme under Uncertain Holding Times

A competitive admission and routing scheme [9] strives to contain loss (4) under the worst case scenario sequence of future request arrivals θ. It was shown [9] that these maximum losses can be bounded by $O(\log v\tau)$ assuming that the holding time τ becomes known upon arrival of the request, and where v is some constant. This bound can be achieved with the cost-based admission control and routing strategy with additive route costs, i.e., the cost c_r of a route r is a sum of the costs c_j of the links j comprising the route r: $c_r = \sum_{j \in r} c_j$. The cost of a link j is a function of the instantaneous load carried by the link x and holding time τ: $c_j = c_j(\tau, x)$. In a case of throughput maximization [9] the link j cost is

$$c_j(\tau, x) = [(2Fh\tau + 1)^{x/B_j} - 1]B_j \quad \textbf{(16)}$$

where F is some constant, the maximum number of hops allowed in a route is h, and link j bandwidth is B_j.

The proposed in this paper framework can be used to mitigate assumption of known, at least probabilistic sense [9], holding times, and address the need to aggregate the real-time information on instantaneous link utilization x. As an illustration consider a case of uncertain holding time $\tau \in [\breve{\tau}, \hat{\tau}]$, and two aggregates: link j is said to be in the aggregate state $y_j = 0$ if the link instantaneous utilization $x_j \in [0, B_{1j}]$ and $y_j = 1$ if $x_j \in [B_{1j}, B_j]$ where $B_{1j} \in [0, B_j]$ is

some threshold. The uncertainties in τ and x_j cause the following uncertainty in the route r cost c_r:

$$\sum_{j \in r} \breve{c}_j(y_j) \equiv \breve{c}_r \le c_r \le \hat{c}_r \equiv \sum_{j \in r} \hat{c}_j(y_j) \tag{17}$$

where $\breve{c}_j(0) = 0$, $\breve{c}_j(1) = c_j(\breve{\tau}, B_{1j})$, $\hat{c}_j(0) = c_j(\hat{\tau}, B_{1j})$, $\hat{c}_j(1) = 2Fh\hat{\tau}B_j$. Under approximation of separable route costs (13), subsection 3.2 results define the aggregated version of the competitive scheme under uncertain holding times.

4.2 Aggregated Minimum Interference Routing Algorithm

Minimum Interference Routing Algorithm (MIRA) [10] does not assume any knowledge of future request arrivals or holding times, but takes advantage of the known network topology by selecting routes that do not "interfere too much" with a route that may be critical to satisfy future demands. The problem was motivated by the need of service providers to set up bandwidth guaranteed path in their backbone or networks. An important context in which these problems arise is that of dynamic Label Switching Path (LPS) set up in Multi-Protocol Label Switched (MPLS) networks. In MPLS packets are encapsulated at ingress points, with labels that are then used to forward the packets along LPSs. Service providers can use virtual circuit switched, bandwidth guaranteed LPSs as component of an IP Virtual Private Network (VPN) service with the bandwidth guarantees used to satisfy customer service-level agreements (SLAs).

MIRA is a heuristic, on-line, state-dependent, cost-based routing algorithm with additive route costs: $c_r = \sum_{j \in r} c_j$. The link j cost c_j is a function of the vector of residual link capacities $z = (z_i)$ in all links i:

$$c_j(z) = \sum_{(n,m)} w_{nm} \delta_{nm}^{(j)}(z) \tag{18}$$

where (n,m) are all possible ingress-egress pairs, w_{nm} is the revenue generated by providing an unit of bandwidth to ingress-egress pair (n,m), and $\delta_{nm}^{(j)}(z) = 1$ if link j is critical for ingress-egress pair (n,m), and $\delta_{nm}^{(j)}(z) = 0$ otherwise. Link j is critical for ingress-egress pair (n,m) if the maximum flow for pair (n,m) decreases whenever the residual capacity of link j decreases, and is not critical otherwise. Paper [10] discusses effective algorithms for calculating sets of critical links, i.e., indicators $\delta_{nm}^{(j)}(z)$, and reports simulation results.

The proposed in this paper framework can be used to mitigate the following problems of MIRA. The first problem is inability to utilize an available incomplete information on the expected loads. The second problem is high sensitivity to the current set of critical links, which is likely to change in the future due to fluctuations in the vector z. The third problem, related to the second one, is need to aggregate the residual capacities. Due to space constraints we briefly address the second and third problems. Assuming that available information on residual link i capacity z_i is characterized by aggregates $z_i \in [\breve{z}_i, \hat{z}_i]$, the range for the link j cost c_j is

$$\sum_{(n,m)} w_{nm} \delta_{nm}^{(j)}(.,\hat{z}_{j-1},\breve{z}_j,\hat{z}_{j+1},.) \le c_j \le \sum_{(n,m)} w_{nm} \delta_{nm}^{(j)}(.,\breve{z}_{j-1},\hat{z}_j,\breve{z}_{j+1},.). \quad \textbf{(19)}$$

Under approximation of separable route costs (13), subsection 3.2 results define the aggregated version of MIRA.

References

1. Courcoubetis, C., Kelly, F.P., and Weber, R., Measurement-based usage charges in communication networks. http://www.statslab.cam.ac.uk/Reports/1997/
2. Bonomi, F., Mitra, D, and Seery, J., Adaptive algorithms for feedback-based flow control in high-speed wide-area networks. *IEEE JSAC*, 13 (1995) 1267-1283
3. Low, S.H., and Varaiya, P.P., A new approach to service provisioning in ATM networks. *IEEE Trans. On Networking*, 1 (1993) 547-553
4. MacKie-Mason, J.K. and Varian, H.R., Pricing congestible network resources. *IEEE JSAC*, 13 (1995) 1141-1149
5. Shenker, S., Fundamental design issues for the future Internet. *IEEE JSAC*, 13 (1995) 1176-1188
6. Gibbens, R. and Kelly, F.P., Resource pricing and evolution of congestion control. http://www.statslab.cam.ac.uk/~frank/PAPERS/evol.html
7. Kelly, F.P., Mauloo, A., and Tan, D., Rate control in communication networks: shadow prices, proportional fairness and stability. J. of the Oper. Res. Soc., 49 (1998) 237-252
8. Mahajan, R., Bellovin, S.M., Floyd, S., Ioannidis, J., Paxson, V., and Shenker, S., Controlling high bandwidth aggregates in the network. http://www.research.att.com/~smb/papers/
9. Plotkin, S., Competitive routing of virtual circuits in ATM networks. *IEEE JSAC*, 13 (1995) 1128-1136
10. Kar, K., Kodialam, M., and Lakshman, T.V., Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications. *IEEE JSAC*, 18 (2000) 2566-2579
11. Harsanyi, J.C., Papers in Game Theory. Reidel Publishing Company (1982)
12. Blackwell, D. and Girschick, M., Theory of Games and Statistical Decisions. Wiley, New York (1954)
13. Marbukh, V., Network management under incomplete information on the operational environment. Intern. Symp. on Inform. Theory and its Appl. (ISITA2000) 637-640